Exponential Distribution and Central Limit Theorem

Liubov Gryaznova

November 21, 2015

Overview

In this project we explore the distribution of averages of 40 exponentials using R to do 1000 simulations. In particular, we compare the sample mean and variance to the theoretical mean and variance of the distribution, and we demonstrate that this distribution of averages is approximately normal in accordance with the Central Limit Theorem even though the random exponentials used to calculate the average are distributed exponentially.

Simulations

The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda (λ) is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation is also $\frac{1}{\lambda}$. For the sake of this assignment we set $\lambda = 0.2$ for all of the simulations. We will investigate the distribution of averages of 40 exponentials, and we will repeat this simulation 1000 times, so we store these values in `n` and `nosim` variables respectively. The result of the simulation is stored in `simulation` vector.

```
lambda <- 0.2         ## rate
n <- 40               ## number of exponentials to take their mean
nosim <- 1000         ## number of simulations
set.seed(1)           ## to improve reproducibility

simulation <- NULL
for (i in 1:nosim) { simulation = c(simulation, mean(rexp(n, lambda))) }
```

Sample Mean versus Theoretical Mean

According to the Central Limit Theorem (CLT) the sample average is approximately normally distributed with a mean given by the population mean and a standard deviation given the standard error of the mean. In other words, the theoretical mean for our distribution is $\frac{1}{\lambda} = 5$. We calculate the sample mean to demonstrate that it is pretty close to the theoretical mean.

```
1/lambda              ## theoretical mean
```

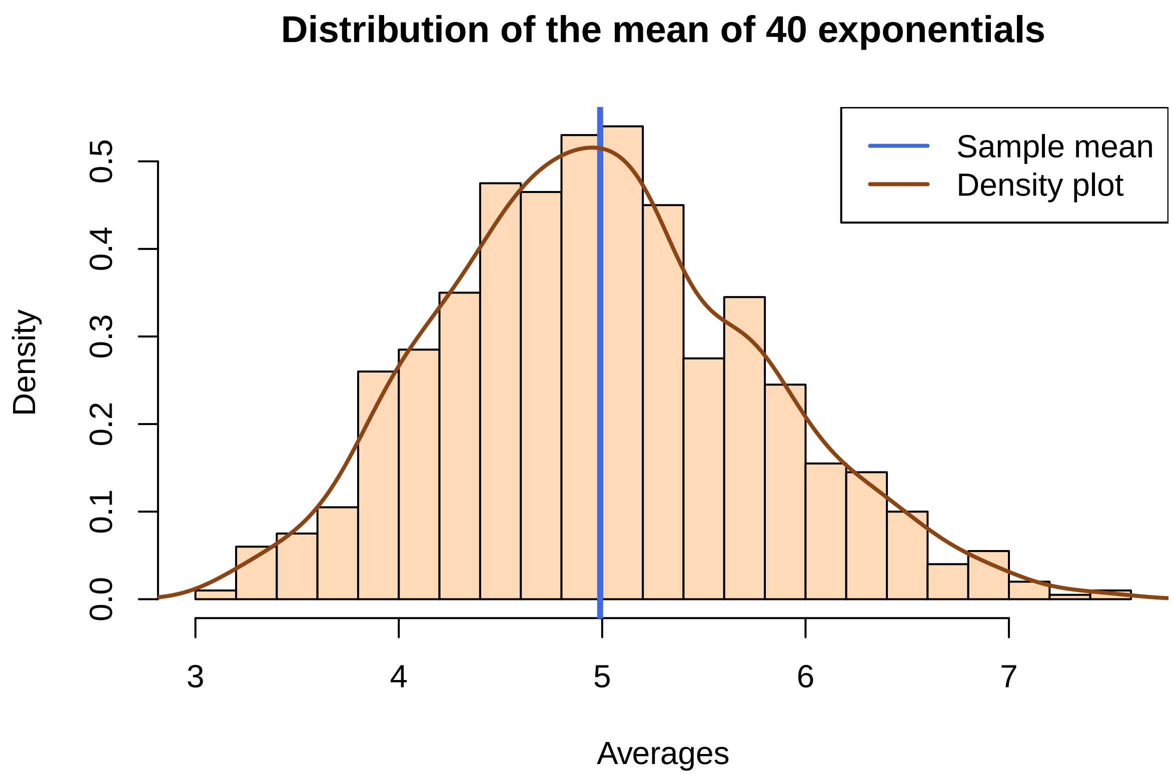
```
## [1] 5
```

```
mean(simulation)      ## sample mean
```

```
## [1] 4.990025
```

The histogram of the sample averages confirms that the distribution is centered at almost the theoretical mean; note that the distribution looks normal.

```r
hist(simulation, probability = TRUE, breaks = 30, col = "peachpuff",
     main = "Distribution of the mean of 40 exponentials", xlab = "Averages")
abline(v = mean(simulation), col = "royalblue", lwd = 3)
lines(density(simulation), lwd = 2, col = "chocolate4")
legend(x = "topright", c("Sample mean", "Density plot"),
       col = c("royalblue", "chocolate4"), lwd = c(2, 2, 2))
```



Sample Variance versus Theoretical Variance

The CLT states that the standard deviation of this distribution is equal to the standard error of the mean: $\sigma/\sqrt{n}$ where σ is the standard deviation of the population and n is the sample size. So the theoretical variance is the theoretical standard deviation squared, or the population variance squared divided by the sample size: σ^2/n. Therefore the theoretical variance of our distribution is $\frac{1}{\lambda^2 \times n}$.

From the calculations below we see that the sample variance 0.611 is pretty close to the theoretical variance 0.625.

```r
(1/lambda)^2 / n        ## theoretical variance
```

```
## [1] 0.625
```

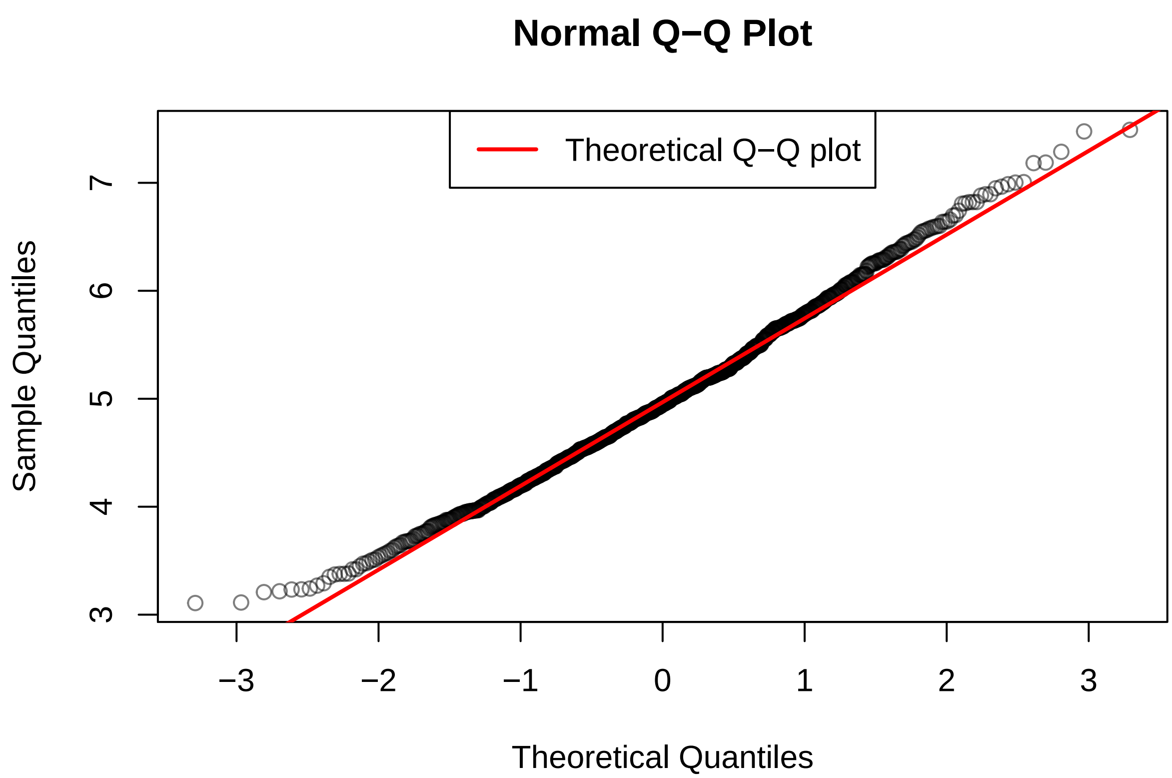
```r
var(simulation)         ## sample variance
```

```
## [1] 0.6111165
```

Distribution

We saw on the histogram above that the distribution of averages of 40 exponential looks Gaussian with the bell-shaped density line. We also confirmed that the sample mean and variance are very close to the theoretical mean and variance. To confirm the normality we make a Q-Q plot.

```
qqnorm(simulation, col = rgb(0, 0, 0, 0.5))
qqline(simulation, col = "red", lwd = 2)
legend(x = "top", "Theoretical Q-Q plot",  col = "red", lwd = c(2, 2, 2))
```



Since the points fall approximately along a red reference line we can conclude that the distribution is close to normal. And we know from the CLT that as the sample size increases the normality increases, too.